# ML FOR CHEMISTRY AND MATERIALS

CyberTraining Workshop and Summer School 2026

Johannes Hachmann

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

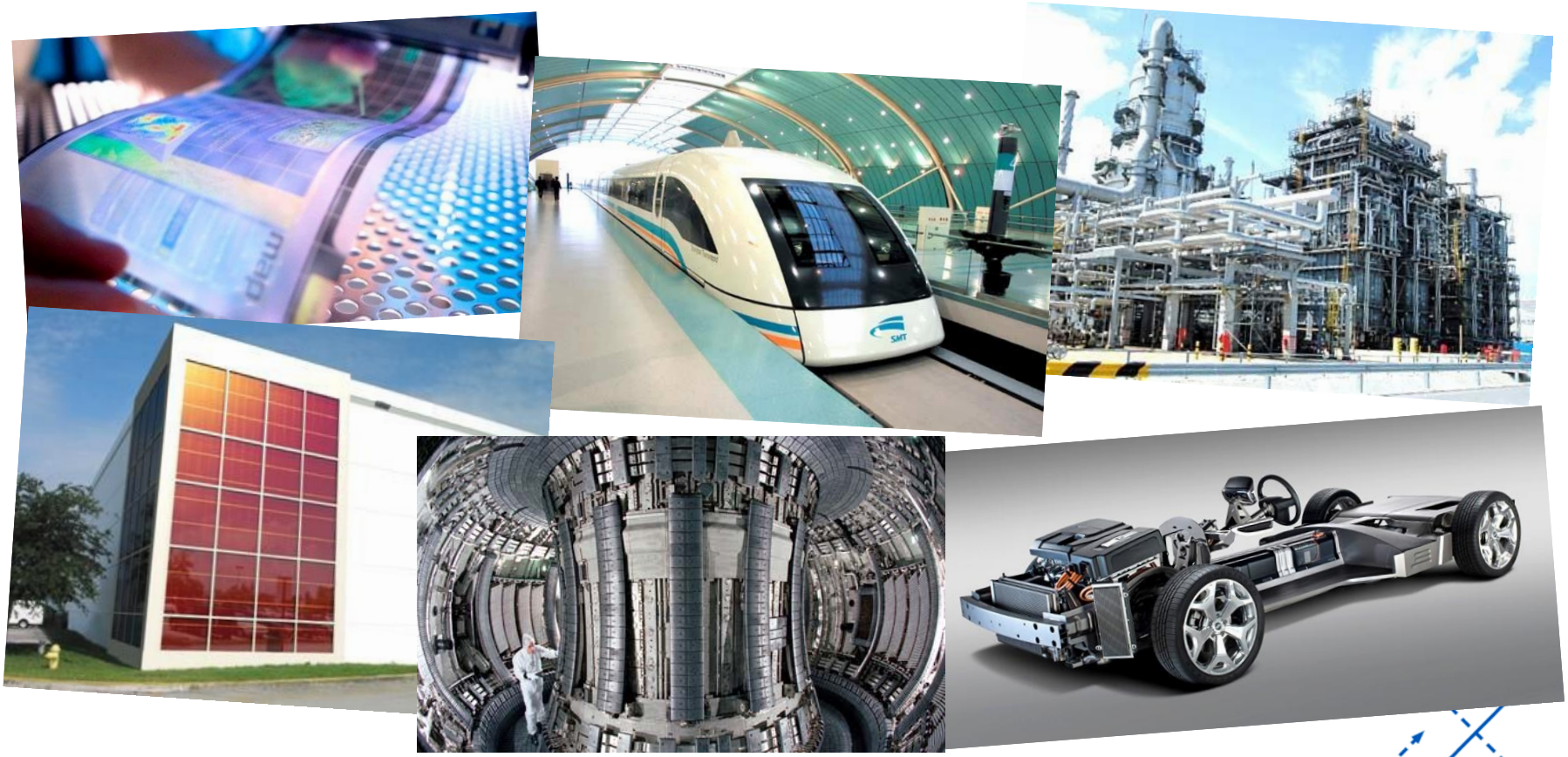
# Acknowledgements

University at Buffalo
The State University of New York

# The role of chemical & materials research

University at Buffalo
The State University of New York

**Real-world solutions** to mankind's **grand challenges** in **innovation, energy, and sustainability** can be found in the discovery and development of **novel compounds, processes, and materials**.

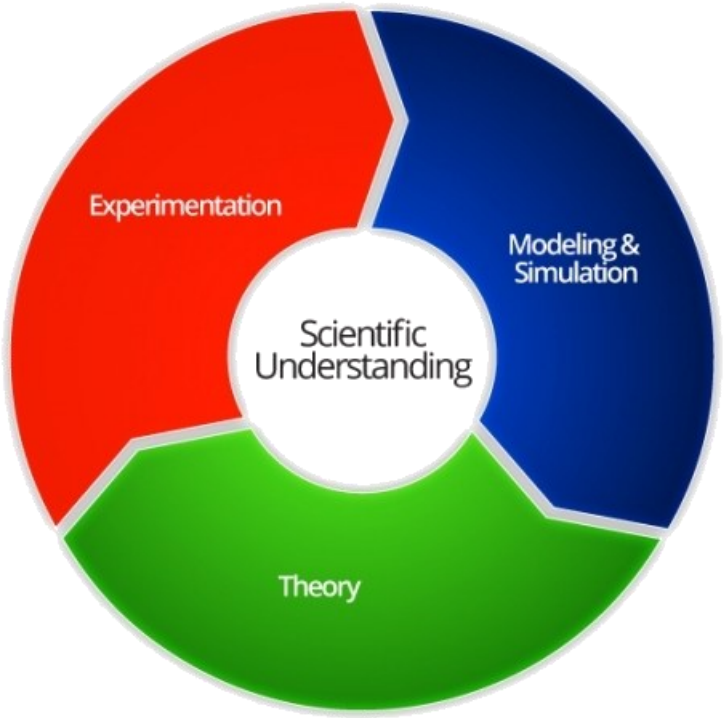


# Theory-driven chemical & materials research

University at Buffalo
The State University of New York

**Real-world solutions** to mankind's **grand challenges** in **innovation, energy, and sustainability** can be found in the discovery and development of **novel compounds, processes, and materials**.

- **progress** increasingly **driven by physics-based modeling, simulation** (QC, MD, ...)




DOE

# Theory-driven chemical & materials research

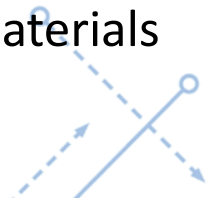
University at Buffalo
The State University of New York

**Real-world solutions** to mankind's **grand challenges** in **innovation, energy, and sustainability** can be found in the discovery and development of **novel compounds, processes, and materials**.

➢ **progress** increasingly **driven by physics-based modeling, simulation** (QC, MD, ...)

- computational studies are **efficient** means to **uncover promising targets** for more time- and resource-intensive investigations in the lab
- they can significantly **accelerate** research by making **guiding predictions**
- they can **provide unique insights** beyond the scope of empirical observation and thus contribute a solid foundation to new findings

➢ <u>thus:</u> they are a valuable asset for **synergistic** research in chemistry, materials

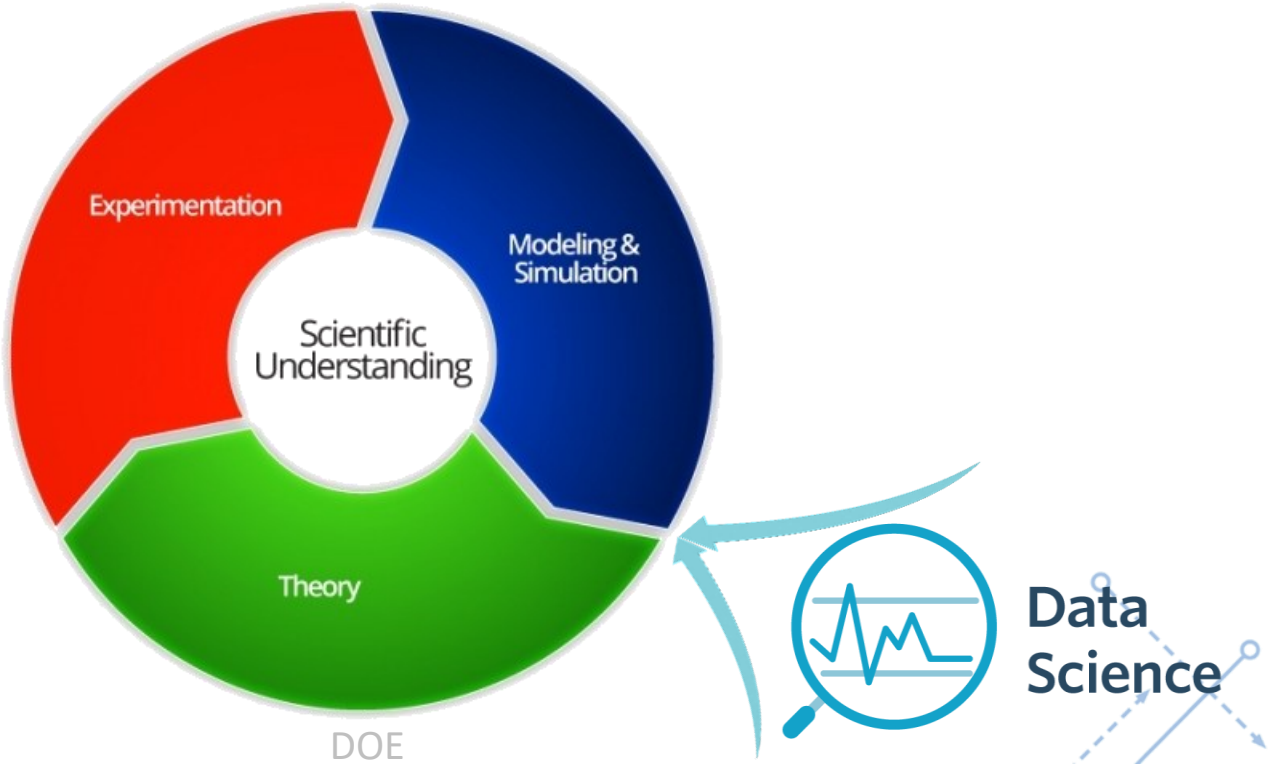


# Data-driven chemical & materials research

University at Buffalo
The State University of New York

**Real-world solutions** to mankind's **grand challenges** in **innovation, energy, and sustainability** can be found in the discovery and development of **novel compounds, processes, and materials**.

- **progress** increasingly **driven by physics-based modeling, simulation** (QC, MD, ...)
- rise of **data-driven** approaches: "*4th pillar of science*"




DOE

# Data-driven chemical & materials research

University at Buffalo
The State University of New York

**Real-world solutions** to mankind's **grand challenges** in **innovation, energy, and sustainability** can be found in the discovery and development of **novel compounds, processes, and materials**.

- **progress** increasingly **driven by physics-based modeling, simulation** (QC, MD, ...)
- rise of **data-driven** approaches: "*4th pillar of science*"
  - VHTS, CI/MI, DM, ML, DL, AI, …

- automated/autonomous VHTS allows for **large-scale *in silico* exploration of chemical space** (its uncharted domains may hold new compound classes with **transformative properties**)
  - generates large data sets

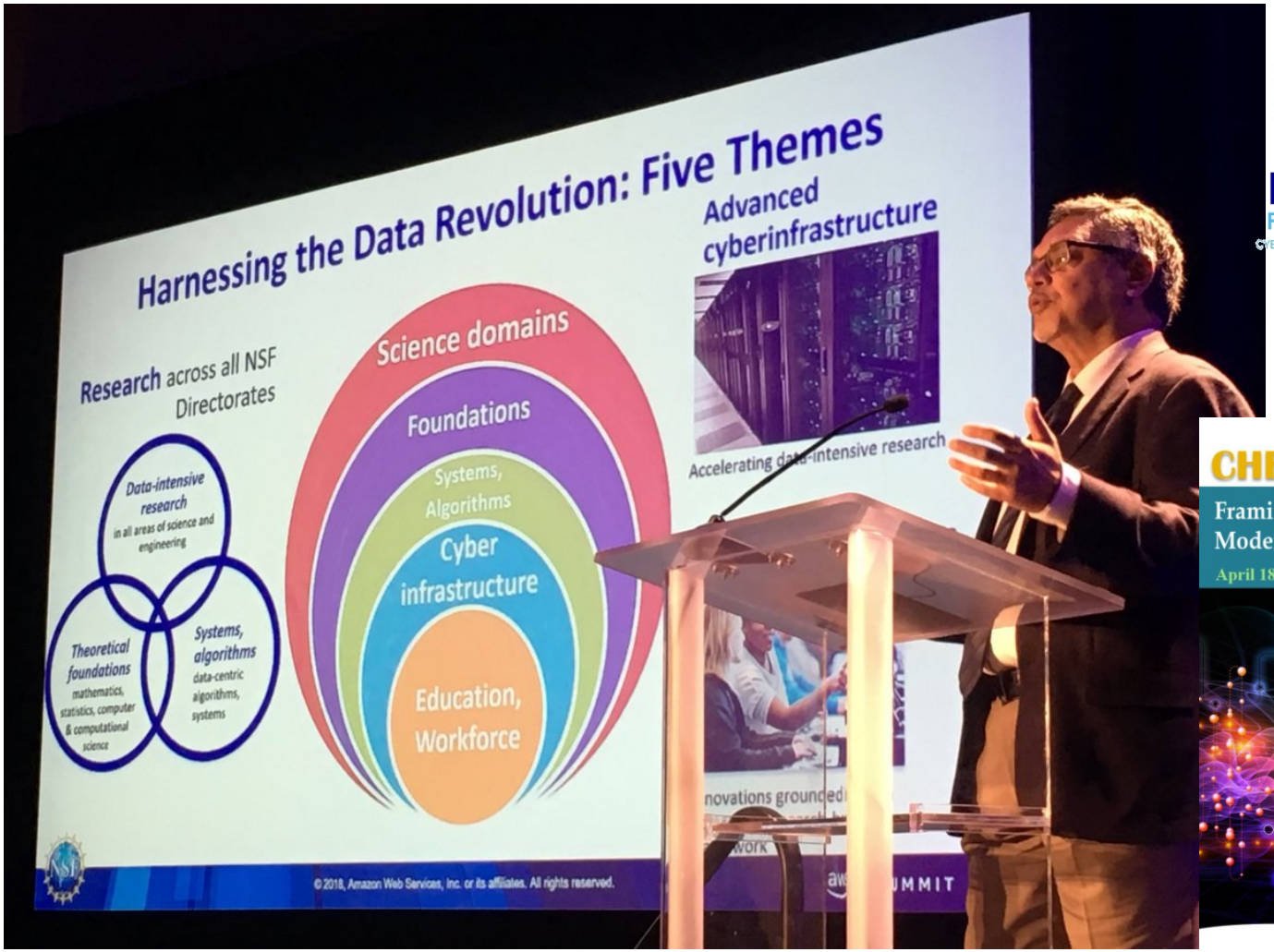
- DM, ML, DL: extract insights, patterns from data sets, create prediction models

# 2011 White House Materials Genome Initiative

University at Buffalo
The State University of New York

# 2017 Data-Driven Discovery Science in Chemistry

University at Buffalo
The State University of New York





# AI *vs* ML *vs* DL

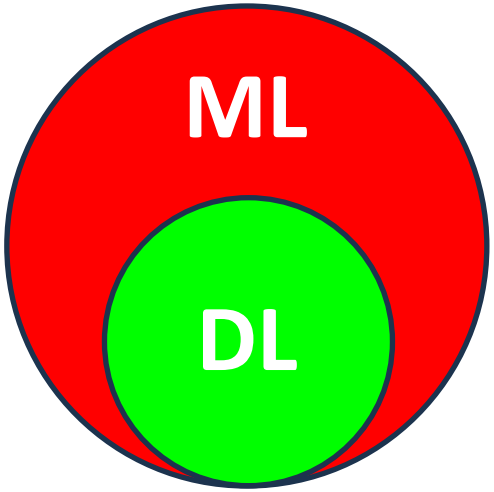
University at Buffalo
The State University of New York

- ML: techniques to generate data-derived prediction models (surrogate models)

# AI *vs* ML *vs* DL

University at Buffalo
The State University of New York

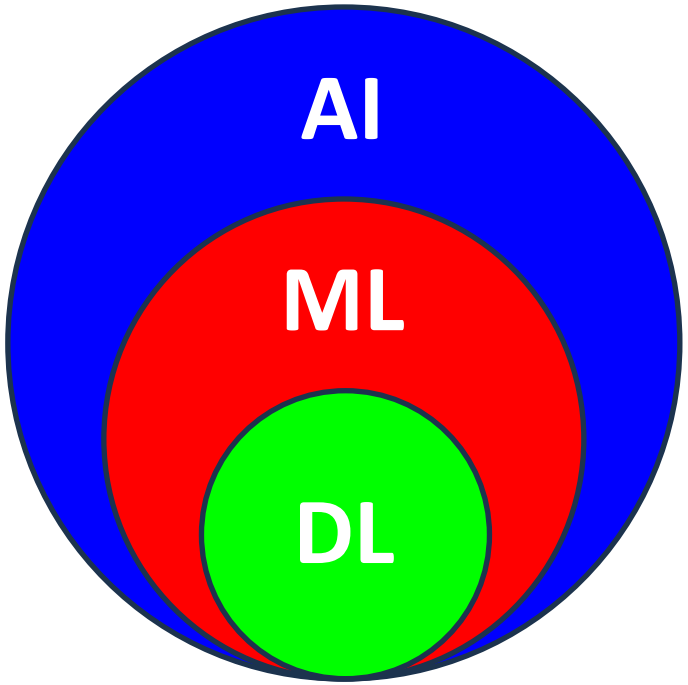
- ML: techniques to generate data-derived prediction models (surrogate models)
- DL: a particular class of ML algorithms (using multi-layer neural networks)



# AI *vs* ML *vs* DL

University at Buffalo
The State University of New York

- ML: techniques to generate data-derived prediction models (surrogate models)
- DL: a particular class of ML algorithms (using multi-layer neural networks)
- AI: techniques to interact with ML models in a human-like fashion



University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

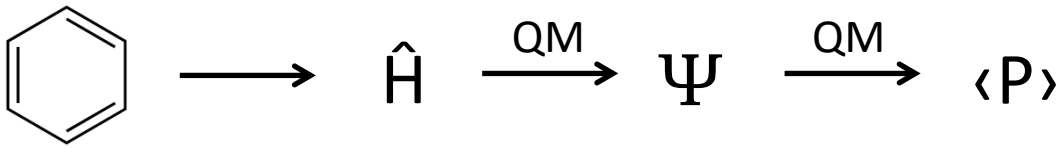
# PHYSICS-BASED *VS* DATA-DERIVED MODELS

What's the difference?

# Physics-based models

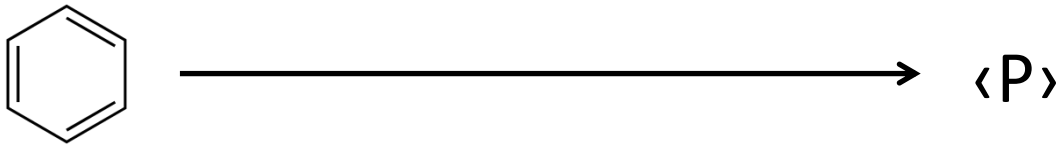
University at Buffalo
The State University of New York

- physics (e.g., QM) rigorously maps a structure to a property




  - i.e., there exists a structure-property causality
  - BUT: QM is **expensive**, we want to recover mapping more efficiently
  - BUT: QM is **opaque**, we want to identify relevant aspects of SPR

  - … could there be a shortcut?



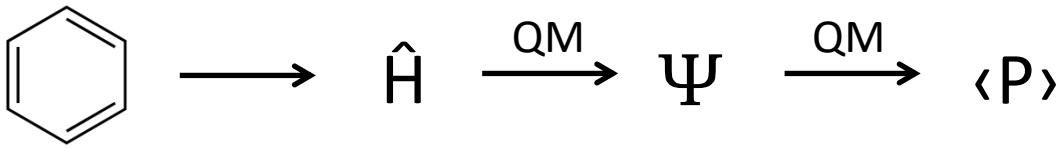



*COChE* 23 (**2019**), 51.

# Data-derived models

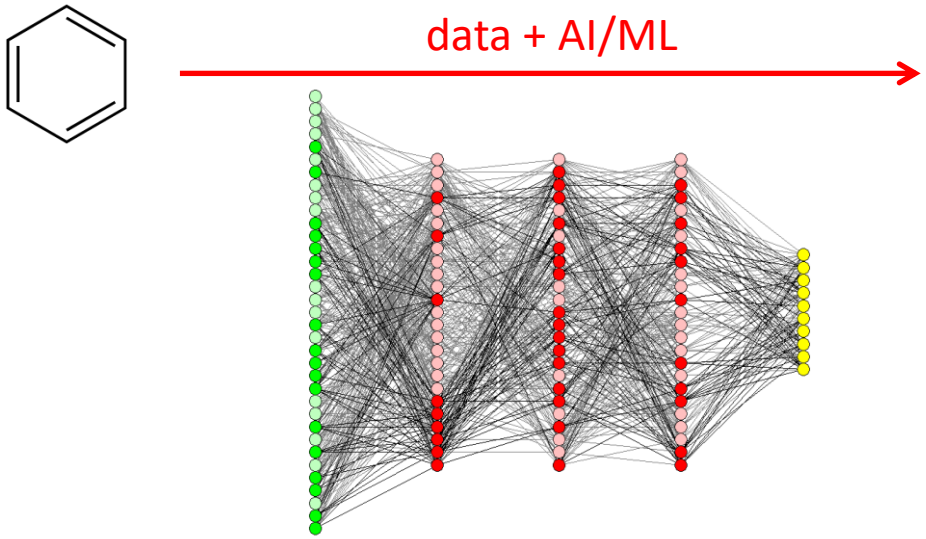
University at Buffalo
The State University of New York

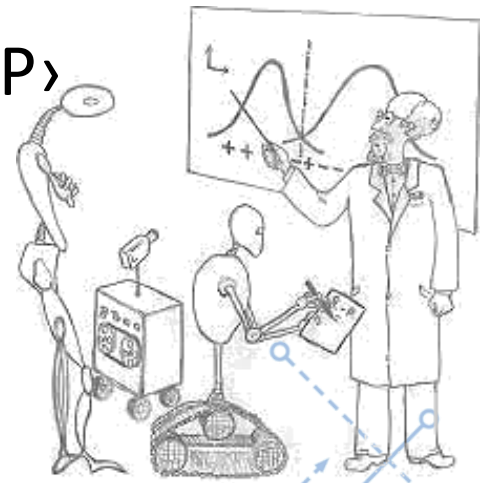
- physics (e.g., QM) rigorously maps a structure to a property




  - i.e., there exists a structure-property causality
  - BUT: QM is **expensive**, we want to recover mapping more efficiently
  - BUT: QM is **opaque**, we want to identify relevant aspects of SPR

  - … could there be a shortcut? *Yes, via AI/ML!*



*COChE* 23 (**2019**), 51.

# Data-derived models

University at Buffalo
The State University of New York

| | property |
|---|---|
| n compounds | … <br> $p_n$ <br> … |

- goal: create prediction model based on set of experimental/computational data



*COChE* 23 (**2019**), 51.

# Data-derived models

University at Buffalo
The State University of New York

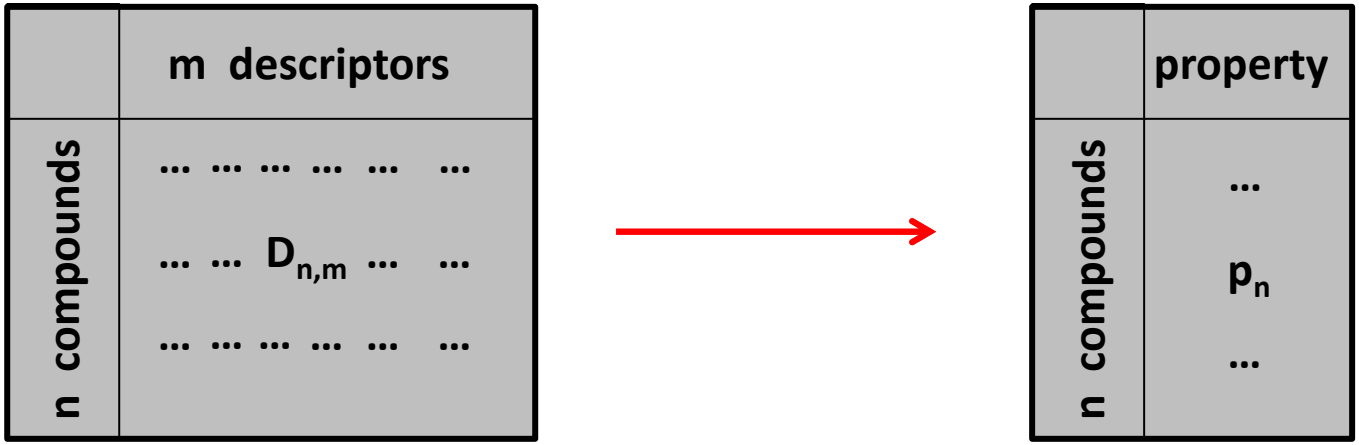

- goal: create prediction model based on set of experimental/computational data



*COChE* 23 (**2019**), 51.

# Data-derived models

University at Buffalo
The State University of New York

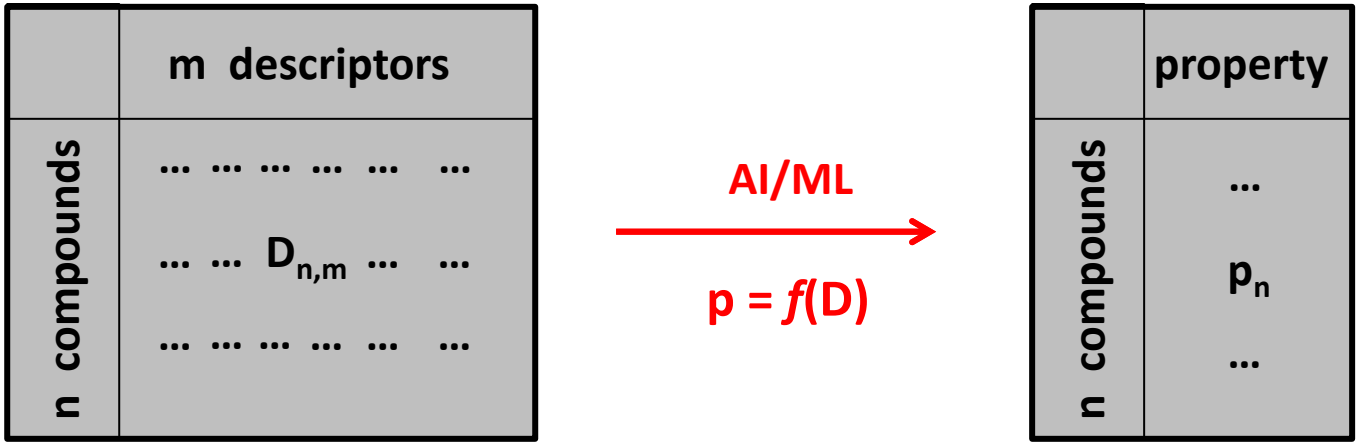



- goal: create prediction model based on set of experimental/computational data



*COChE* 23 (**2019**), 51.

# Data-derived models




- goal: create prediction model based on set of experimental/computational data

- ***f*** is **predictive**, encapsulates essence of **structure-property relationships!**
- we can use ***f*** for **HTP discovery,** $f^{-1}$ for **rational design, inverse engineering!**

- ***f*** is easy to evaluate, allowing for **fast results!**



*COChE* 23 (**2019**), 51.

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

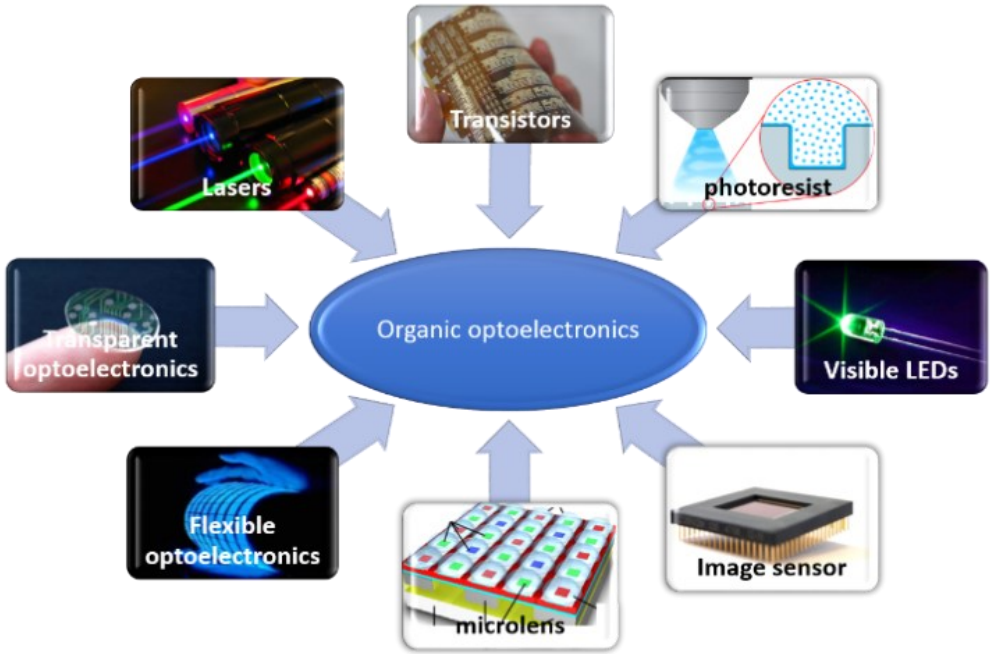
# A WORKING EXAMPLE

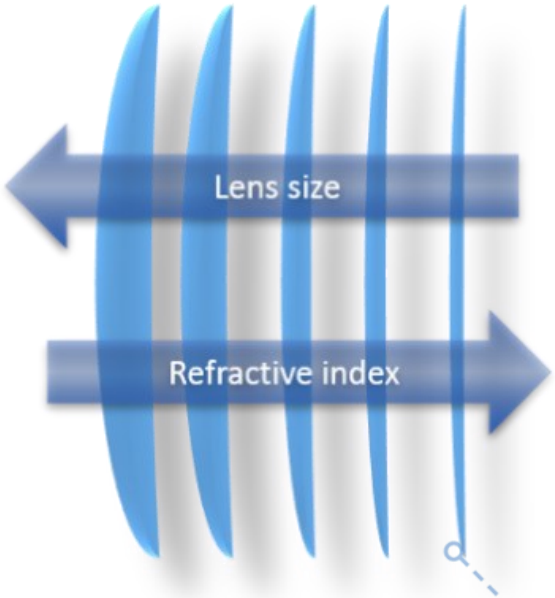
Discovery and design of
high-refractive index polymers

# Ex: Development of high-refractive-index polymers

University at Buffalo
The State University of New York

- plastic material for **optic, optoelectronic applications**:
  - anti-reflective coatings, optical adhesives, substrates for advanced displays, micro-lenses in image sensors
  - *in situ* application
- <u>problem:</u> RI > 1.7 needed, but typical C-based polymers have RI ≈ 1.3 to 1.5





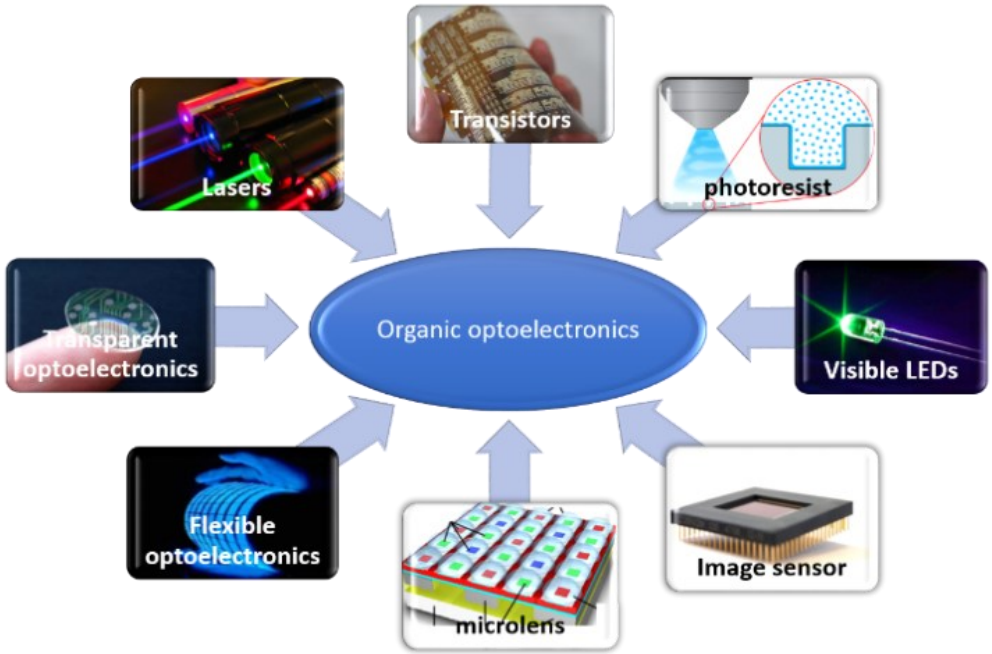

*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

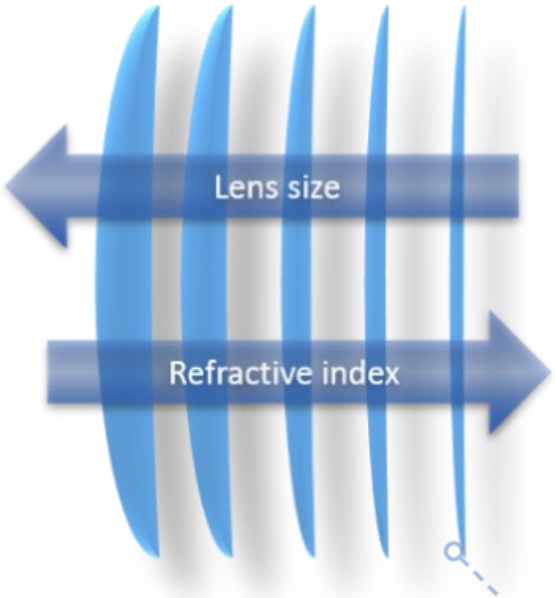
# Ex: Development of high-refractive-index polymers

University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\boldsymbol{\alpha}\,\mathbf{N}/3\epsilon_0}{1 - \boldsymbol{\alpha}\,\mathbf{N}/3\epsilon_0}\right)^{1/2}$$






*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

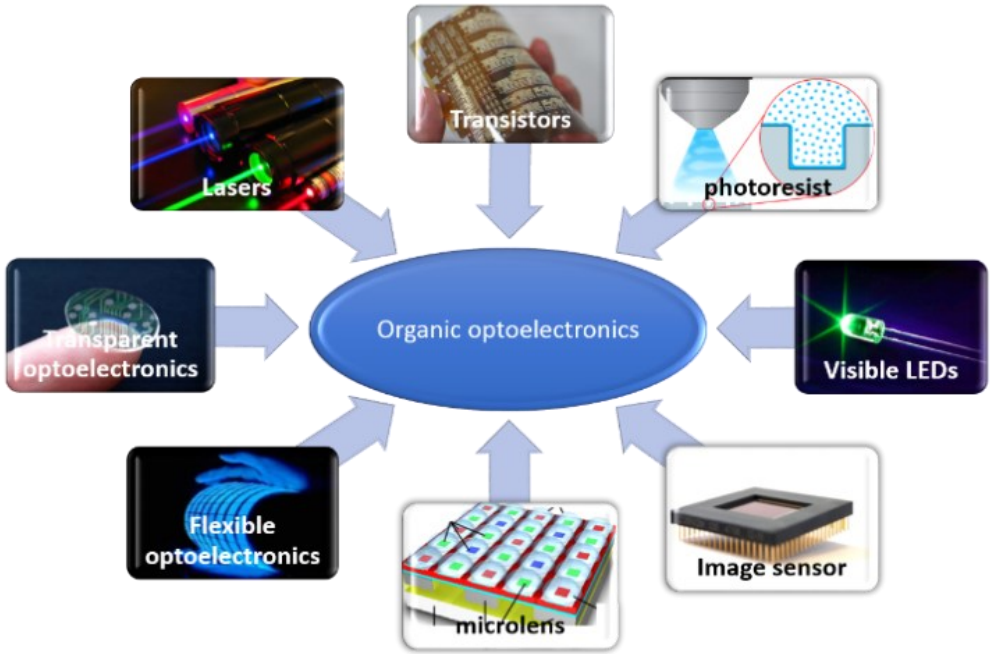
# Ex: Development of high-refractive-index polymers

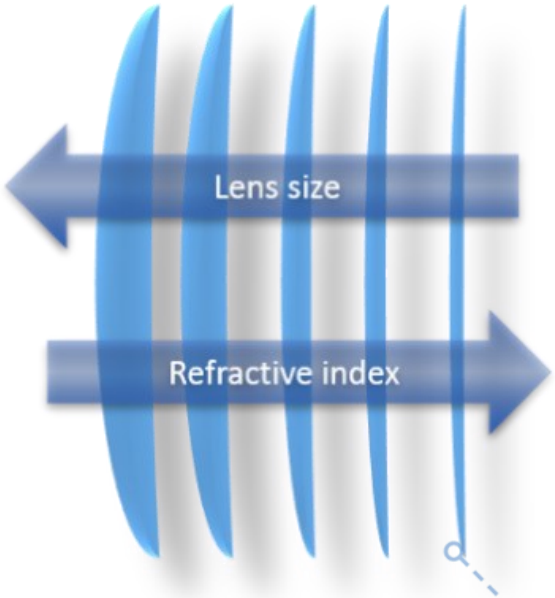
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0}\right)^{1/2}$$

refractive index — molecular polarizability — number density






*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

# Ex: Development of high-refractive-index polymers

University at Buffalo
The State University of New York

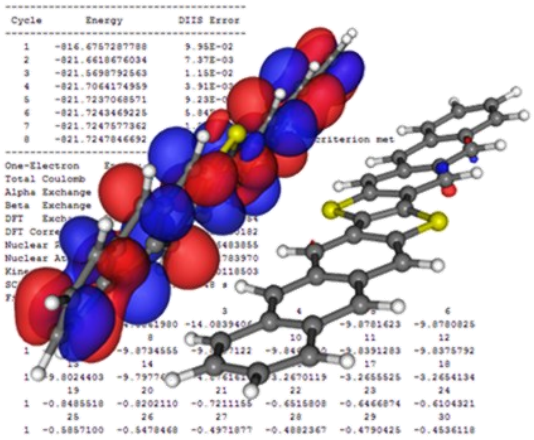
**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

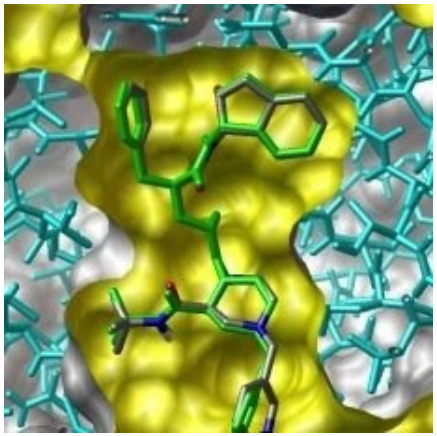
$$n_r = \left(\frac{1 + 2\alpha\,\mathbf{N}/3\epsilon_0}{1 - \alpha\,\mathbf{N}/3\epsilon_0}\right)^{1/2}$$

**refractive index**

**molecular polarizability (e.g., from DFT)**

**number density (e.g., from MD)**







*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

# A WORKING EXAMPLE

Replace physics-based
with data-derived models

# Ex: Development of high-refractive-index polymers

University at Buffalo
The State University of New York

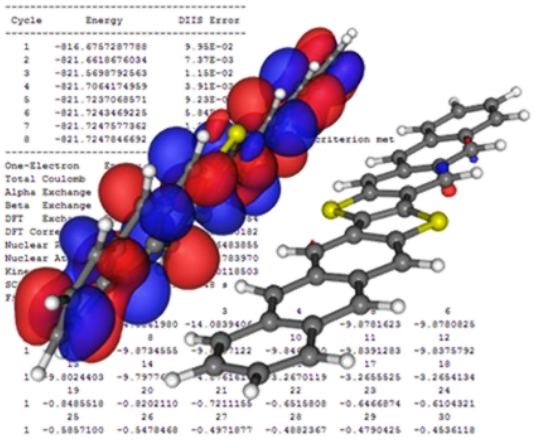
**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

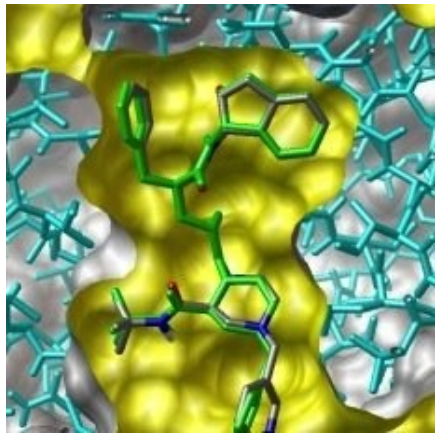
$$n_r = \left(\frac{1 + 2\alpha\,\mathbf{N}/3\epsilon_0}{1 - \alpha\,\mathbf{N}/3\epsilon_0}\right)^{1/2}$$

refractive index

molecular polarizability (e.g., from DFT)

number density (e.g., from MD)







*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

# Ex: Development of high-refractive-index polymers

University at Buffalo
The State University of New York

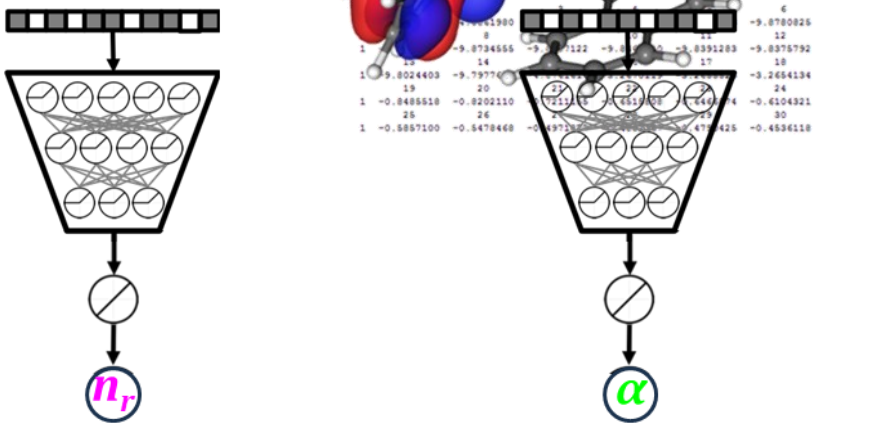
**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

refractive index

molecular polarizability
(e.g., from DFT/ML)

number density
(e.g., from MD/ML)






*JCP* 148 (**2018**), 241712; *JPCC* 123 (**2019**), 14610; *PCCP* 21 (**2019**), 4452.

# Ex: MD number density

University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha \mathbf{N}/3\epsilon_0}{1 - \alpha \mathbf{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \frac{\rho\, M}{N_A}$



*Chem Sci* 10 (**2019**), 8374.

# Ex: MD number density

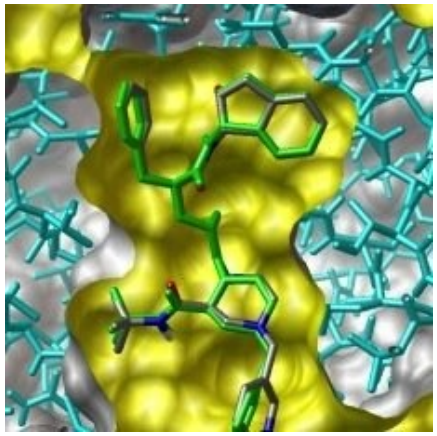
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \dfrac{\rho\, M}{N_A}$
  - MD simulation *via* GROMACS
    - difficult and expensive to compute





*Chem Sci* 10 (**2019**), 8374.

# Ex: MD number density

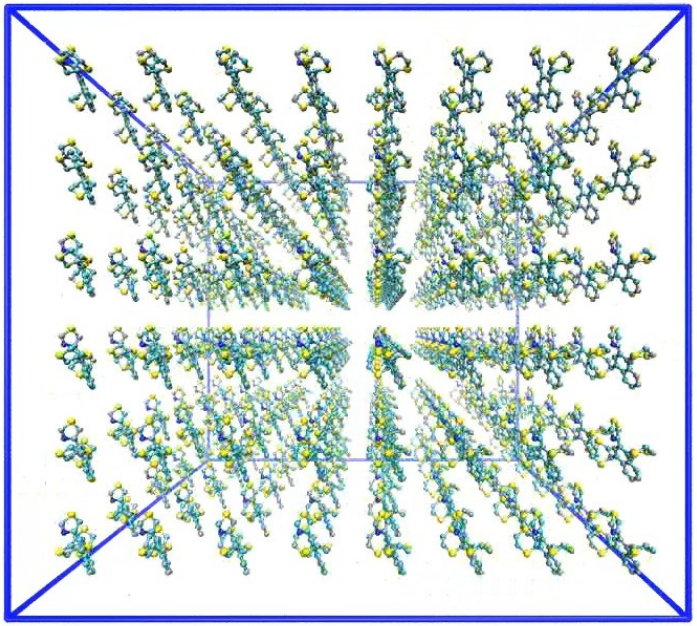
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \dfrac{\rho\, M}{N_A}$
  - MD simulation *via* GROMACS
    - difficult and expensive to compute



NVT step

*Chem Sci* 10 (**2019**), 8374.

# Ex: MD number density

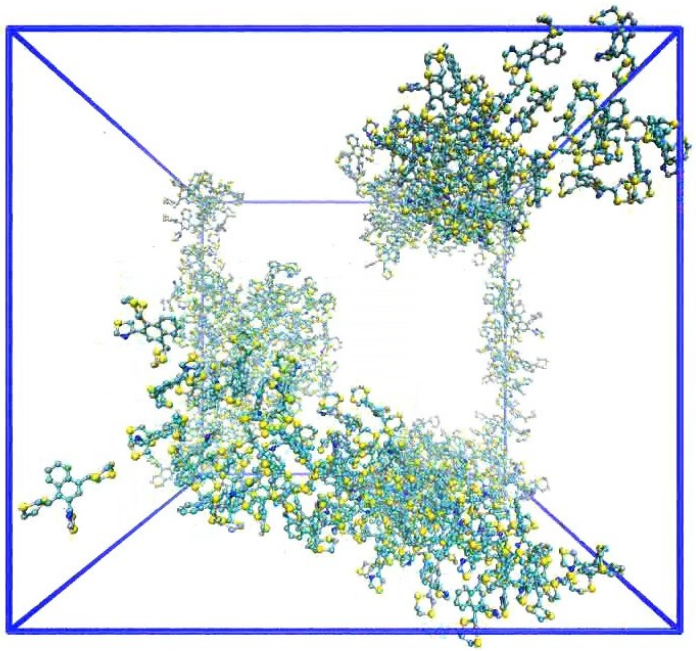
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \dfrac{\rho\, M}{N_A}$
  - MD simulation *via* GROMACS
    - difficult and expensive to compute



NPT step

*Chem Sci* 10 (**2019**), 8374.

# Ex: MD number density

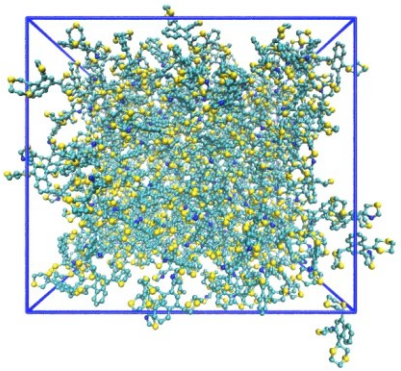
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \dfrac{\rho\, M}{N_A}$
  - MD simulation *via* GROMACS
    - difficult and expensive to compute



*Chem Sci* 10 (**2019**), 8374.

# Ex: MD number density

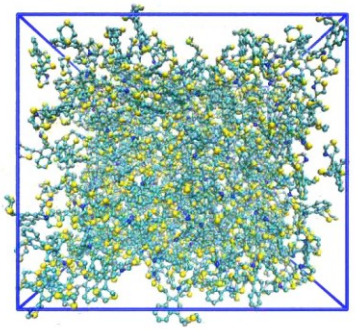
University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha\,\mathrm{N}/3\epsilon_0}{1 - \alpha\,\mathrm{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \dfrac{\rho\, M}{N_A}$
  - MD simulation *via* GROMACS
    - difficult and expensive to compute



*Chem Sci* 10 (**2019**), 8374.

# Ex: ML model for the MD number density

University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

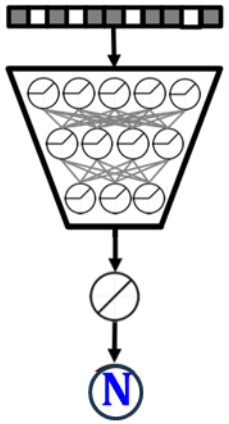
$$n_r = \left(\frac{1 + 2\alpha \mathbf{N}/3\epsilon_0}{1 - \alpha \mathbf{N}/3\epsilon_0}\right)^{1/2}$$

**number density**

- **number density**: $\mathrm{N} = \frac{\rho\, M}{N_A}$

  ~~➢ MD simulation *via* GROMACS~~

  ~~➢ difficult and expensive to compute~~




➢ develop an **ML model** for MD results
- DNN regression model

➢ $R^2$ = 0.98, MAPE = 0.9%, RMSPE = 1.1%; i.e., **within intrinsic MD error**

➢ **7 orders of magnitude faster than MD**



*ChemSci* 10 (**2019**), 8374.

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

# A WORKING EXAMPLE

Correct physics-based models with data-derived corrections

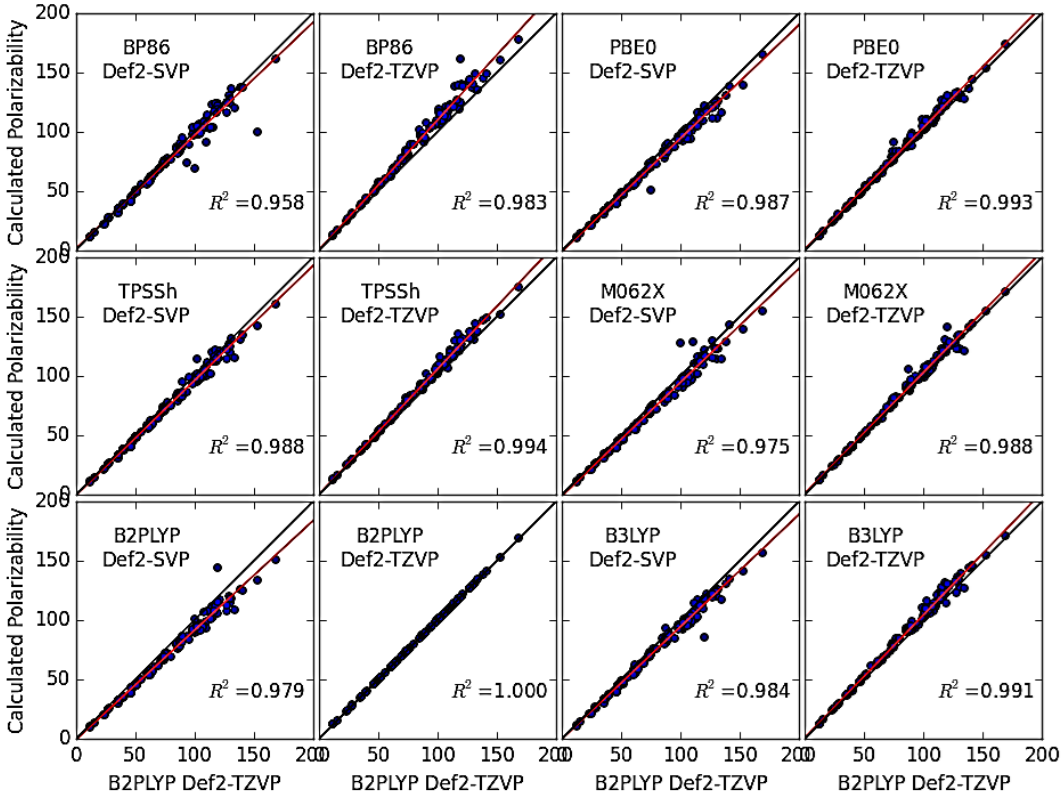
# Ex: DFT polarizabilities

University at Buffalo
The State University of New York

**Lorentz-Lorenz equation**:
(Clausius-Mossotti relation)

$$n_r = \left(\frac{1 + 2\alpha \mathbf{N}/3\epsilon_0}{1 - \alpha \mathbf{N}/3\epsilon_0}\right)^{1/2}$$

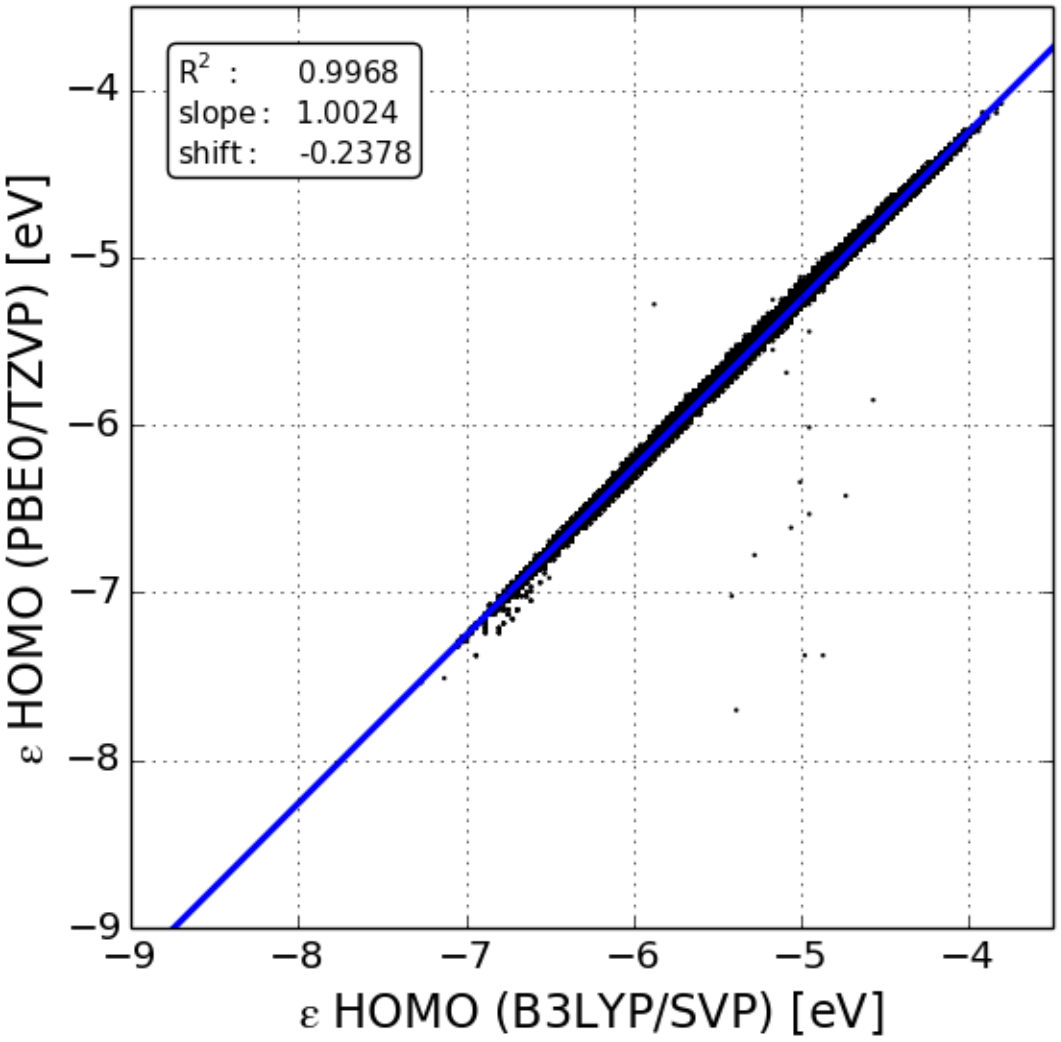
molecular polarizability




*PCCP* 21 (**2019**), 4452.

# Ex: Correcting systematic errors in DFT results



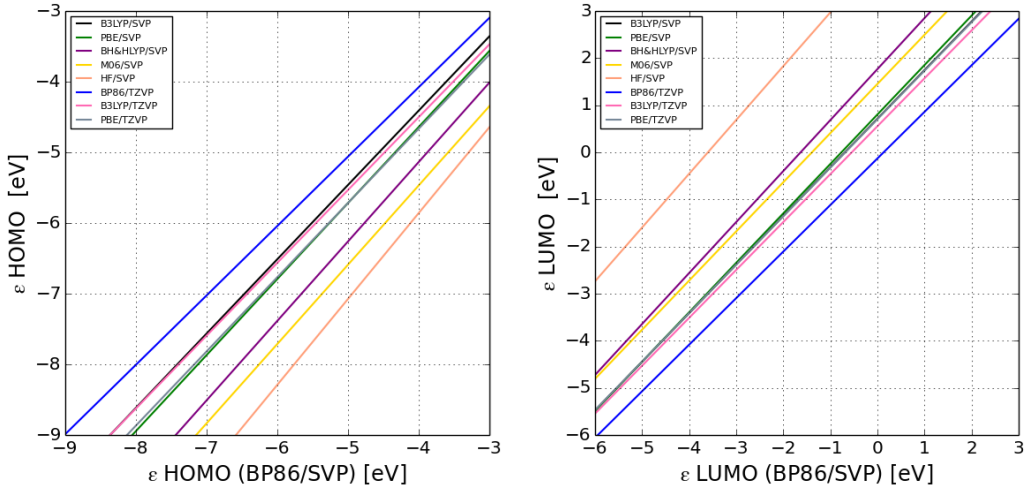



- DZ-hybrid *vs* TZ-hybrid: excellent correlation; straightforward projection

# Ex: Correcting systematic errors in DFT results

University at Buffalo
The State University of New York

- different physics-based methods yield different data
  - typically, better data is more expensive
  - BUT different data sets are systematically correlated
    - ➢ **exploit by projecting cheap onto expensive data** (e.g., via Δ-learning or transfer learning)

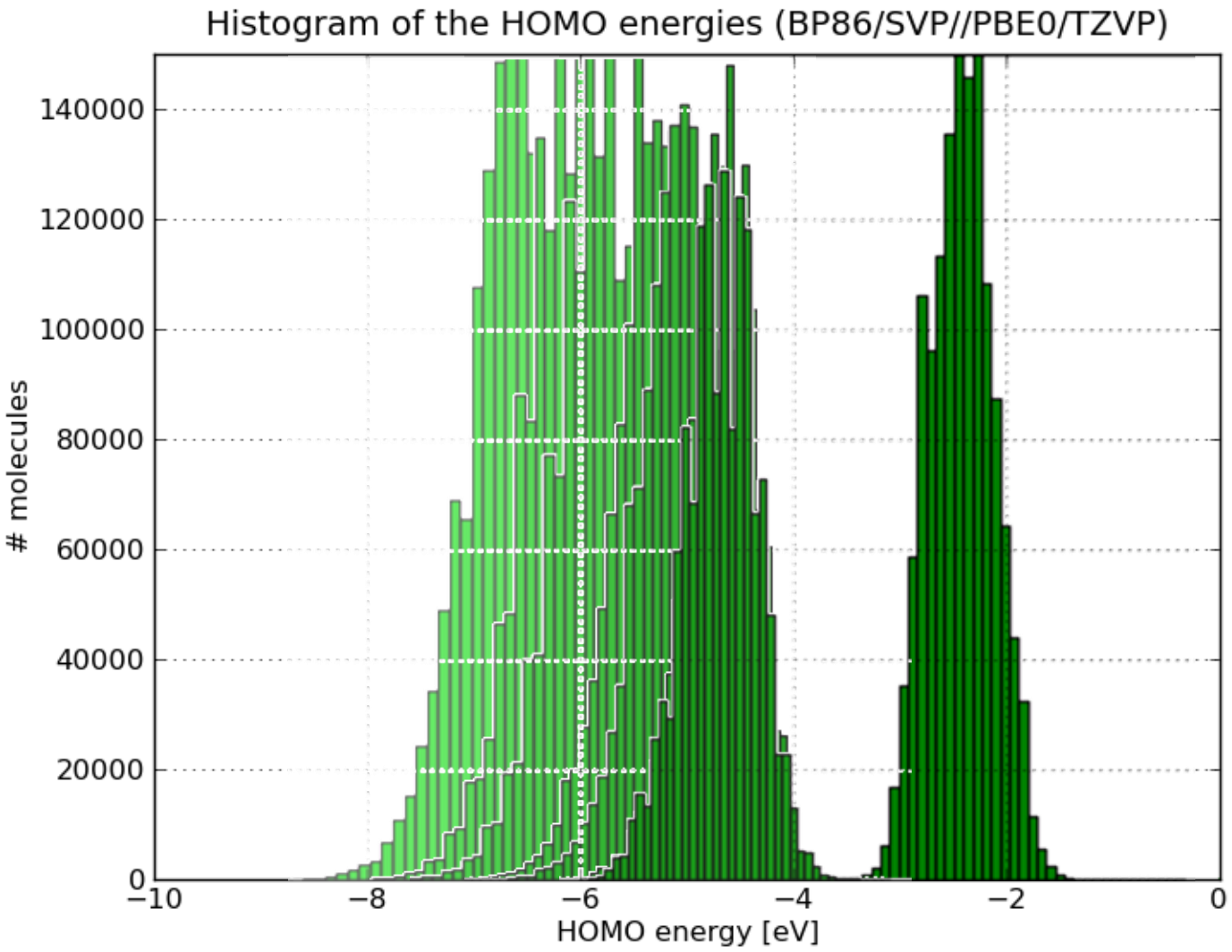





*PCCP* 21 (**2019**), 4452.

# Ex: Correcting systematic errors in DFT results

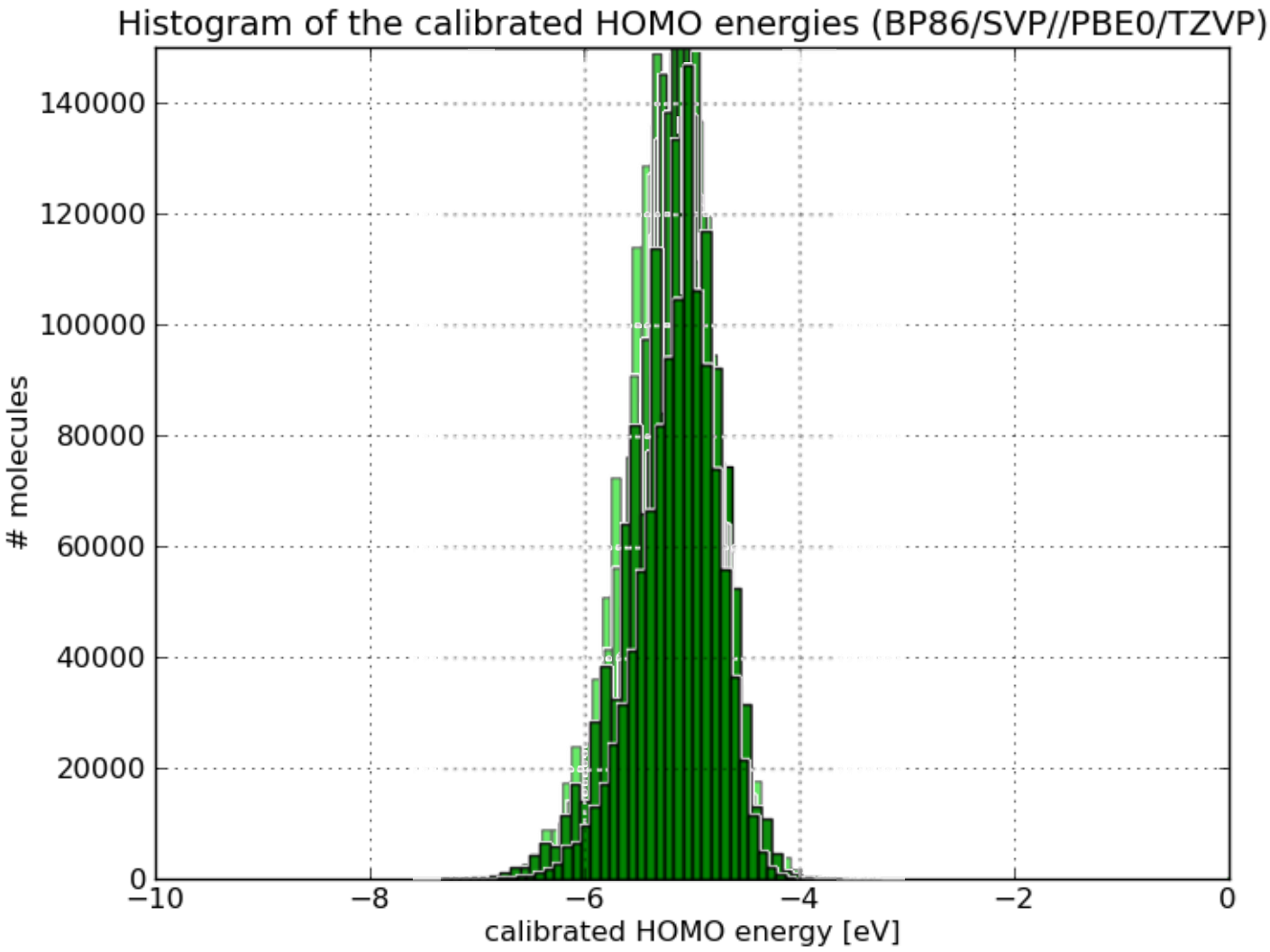
University at Buffalo
The State University of New York




*EES* **7** (**2014**), 698.

# Ex: Correcting systematic errors in DFT results

University at Buffalo
The State University of New York




*EES* **7** (**2014**), 698.

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

# SO…
# HOW DO WE DO IT?

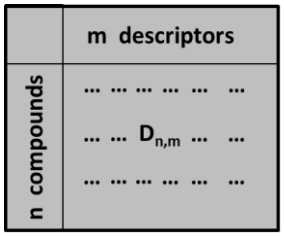
How to make AI/ML work in the chemistry & materials domain

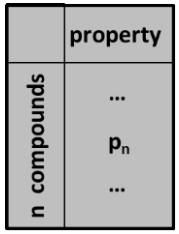
# How to make AI/ML work in chemistry & materials

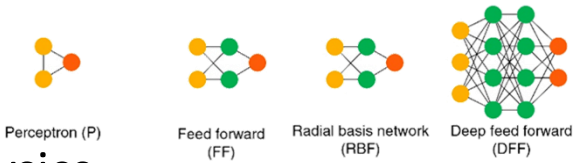
University at Buffalo
The State University of New York

**Challenges for a given problem (… and for the field as a whole)**

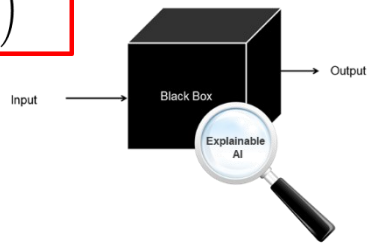
- adequate featurization
  - captures its inherent physics *and* is accessible/affordable
- access to training data of sufficient quality and quantity
  - alternatively: AI/ML for 'small' and/or heteroscedastic data
- identify viable AI/ML approaches
- leverage/incorporate some of the known physics
- explain resulting models and predictions
  - open AI/ML 'black-box', address correlation *vs* causation
- ensure that virtual results are meaningful in the real world
  - e.g., virtually proposed compounds have to be synthetically accessible
  - e.g., AI/ML applicability for extreme value predictions
- realize generative inverse design ideas based on SPR insights
- provide access to all this to community at large

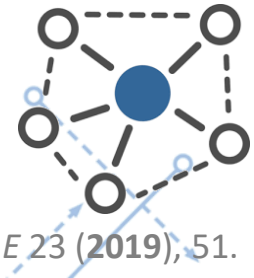



$$n_r = \left(\frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0}\right)^{1/2}$$





*COChE* 23 (**2019**), 51.

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

# GETTING STARTED WITH ML IN S&E

Important processes and workflows in ML

University at Buffalo
Department of Chemical and Biological Engineering
School of Engineering and Applied Sciences

# GETTING STARTED WITH ML IN S&E

Python and its important libraries

# Our software ecosystem for data-driven research

University at Buffalo
The State University of New York


ChemML
0.4.3

Search docs

CHEMML WRAPPER DOCUMENTATION
ChemML Wrapper Tutorial
Input File Manual
Input File GUI
Table of Contents
Wrapper Reference
CHEMML LIBRARY DOCUMENTATION
cheml library



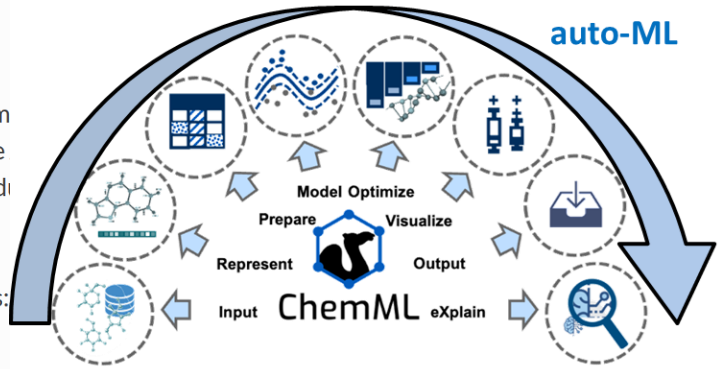

## Welcome to the ChemML's documentation!

ChemML is a machine learning and informatics program suite for the analysis, mining, and modeling of chemical and materials data.

- source: https://github.com/hachmannlab/chemml
- documentation: https://hachmannlab.github.io/chemml

## Code Design:

ChemML is developed in the Python 2 programm
analysis and ML libraries (accessible through the
libraries. The development follows a strictly mod
code as flexible and versatile as possible.




The package consists of two python frameworks:

- **ChemML library (cheml):**

  It is a host for all the methods that are developed or coded from scratch by developers. The format of library is similar to the Scikit-learn library.



*MolSim* 44 (**2018**), 921; *WIRE CMS* (**2020**), e1458; https://github.com/hachmannlab/chemml

# Take-home messages

University at Buffalo
The State University of New York

AI/ML can become a mainstream tool in chemistry & materials research by…

- … finding patterns in complex data, synthesizing insights;
- … complementing experiment and physical-based models;
- … augmenting, calibrating, or otherwise correcting results;
- … incorporating known physics into data-derived models;
- … guiding studies and their objectives;
- … deepening our understanding of structure-property relationships;
- … accelerating discovery and enabling inverse *de novo* design.


University at Buffalo

# Department of Chemical and Biological Engineering

School of Engineering and Applied Sciences